DOCUMENT RESUME

ED 212 664 TM 820 073

AUTHOR Wilcox, Rand R.
TITLE A Lower Bound to the Probability of Choosing the Optimal Passing Score for a Mastery Test When There is an External Criterion [and] Estimating the Parameters of the Beta-Binomial Distribution.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Nov 78
GRANT OB-NIE-G-78-0213
NOTE 33p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Cutting Scores; Factor Analysis; Mastery Tests; Mathematical Models; *Maximum Likelihood Statistics; Probability; *Scoring Formulas
IDENTIFIERS Beta Binomial Test Model; Estimation

ABSTRACT

A mastery test is frequently described as follows: an examinee responds to n dichotomously scored test items. Depending upon the examinee's observed (number correct) score, a mastery decision is made and the examinee is advanced to the next level of instruction. Otherwise, a nonmastery decision is made and the examinee is given remedial work. This document deals with the problem of determining an optimal passing score for a mastery test when the purpose of the test is to predict success or failure on an external criterion. For the case of constant losses for the two error types, a method of determining an optimal passing score is readily derived using standard techniques. The purpose of this research is to describe a lower bound to the probability of identifying an optimal passing score based on a random sample of examinees. The second section of this document deals with the necessity of estimating two parameters to approximate the maximum liklihood estimates using iterative technique. Using Monte Carlo techniques, the Newton-Raphson approximation is compared to other procedures. (Author/CE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED212664

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

DELIVERABLE: November 1978

STUDIES IN MEASUREMENT AND METHODOLOGY

Work Unit 3:

Psychometric Problems in Achievement Tests

Rand R. Wilcox
Project Director

A LOWER BOUND TO THE PROBABILITY OF CHOOSING THE OPTIMAL PASSING SCORE FOR A MASTERY TEST WHEN THERE IS AN EXTERNAL CRITERION

Rand R. Wilcox

ESTIMATING THE PARAMETERS OF THE BETA-BINOMIAL DISTRIBUTION

Rand R. Wilcox

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. S. Gray
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Grant Number:
OB-NIEG-78-0213

Center for the Study of Evaluation
Graduate School of Education
University of California - Los Angeles

November 1978

TM 820 073

The work upon which this publication is based was performed pursuant to a grant with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official NIE position or policy.

A LOWER BOUND TO THE PROBABILITY OF CHOOSING THE OPTIMAL PASSING SCORE FOR A MASTERY TEST WHEN THERE IS AN EXTERNAL CRITERION

Rand R. Wilcox
Center for the Study of Evaluation
University of California at Los Angeles

ABSTRACT

Recently there has been interest in the problem of determining an optimal passing score for a mastery test when the purpose of the test is to predict success or failure on an external criterion. For the case of constant losses for the two error types, a method of determining an optimal passing score is readily derived using standard techniques. The purpose of this note is to describe a lower bound to the probability of identifying an optimal passing score based on a random sample of N examinees.

1. Introduction

A mastery test is frequently described as follows: An examinee responds to n dichotomously scored test items. If x, the examinee's observed (number correct) score, is greater than or equal to x_0, a mastery decision is made and the examinee is advanced to the next level of instruction. Otherwise a nonmastery decision is made and the examinee is given remedial work.

One of the fundamental problems associated with mastery tests is determining x_0, the passing score, so that it has certain optimal properties. Huynh (1976, 1977) has considered this problem for the case when an external criterion exists. Let the random variable y equal one or zero according to whether an examinee succeeds or fails at some later task which he/she attempts after taking the mastery test. Within the framework described by Huynh, we predict success (y=1) if $x \geq x_0$ and failure (y=0) if $x < x_0$. Let θ represent the percent correct true score of an examinee. (See Lord and Novick, 1968, Chapter 23.) To implement the general solution he develops to determine an optimal passing score, Huynh assumes that the true scores of the examinees follow a beta distribution and that there exists a function of the true scores, $s(\theta)$, which determines the probability of success on the external criterion. For a specific examinee with true score θ he also assumes that the distribution of observed scores is given by the binomial probability function. Based on these assumptions he first determines an optimal criterion score, θ_0. If $\theta \geq \theta_0$ he predicts success and if $\theta < \theta_0$ he predicts failure. Once θ_0 has been determined, it is possible, at least for some situations, to determine an x_0 that has certain optimal properties. (See also Van der Linden and Mellenbergh, 1977.)

There is evidence that the beta-binomial model might give a reasonably good fit to mental test data (Keats and Lord, 1962). However, there are several reasons why alternative solutions need to be considered. First, there are estimation problems associated with the model. In particular, the estimates of the parameters of the assumed beta prior used by Huynh might yield inadmissible (negative) values in which case Huynh's solution breaks down. Second, even for a moderately large number of examinees, the estimates of these parameters might be wildly inaccurate. It is possible to approximate the maximum likelihood estimates of the parameters of the beta prior using the procedure outlined by Griffiths (1973) but results reported by Wilcox (1979) imply that a large number of examinees might still be required in order to obtain reasonably accurate results. We note that Griffiths' estimation procedure does not eliminate the first difficulty described above. Third, excluding U-shaped distributions of true scores the beta-binomial model has the theoretical difficulty of allowing at most one mode for the true score distribution. Fourth, for the reasons given by Lord (1965), Lord and Novick (1968, Chapter 23) and Meredith and Kearns (1973) the assumption that the conditional distribution of observed scores is given by

$$f(x \mid \theta) = \binom{n}{x} \theta^{x} (1 - \theta)^{n-x},$$

the binomial probability function, must be viewed as an oversimplification of the "true" situation when an item sampling model applies. These last two difficulties are the main motivation for this paper.

In this note, a procedure for determining the optimal passing score is described which makes no assumption about the conditional distribution of

observed scores or the distribution of true scores. The appeal of this procedure is that it is possible to make statements about how certain we can be that an optimal passing score is chosen given the test score results on N examinees. When using Huynh's solution, it is not at all clear how certain we can be that an optimal passing score has been determined.

2. Procedure for Determining the Optimal Passing Score

We assume that there are test results on N randomly chosen examinees. The results are represented by the pair (x_i, y_i) where, as indicated earlier, x_i is the number-correct score on the mastery test and y_i denotes success (y_i=1) or failure (y_i=0) on the external criterion for the ith examinee. Let x_{0j} (j=1,...,k) be the k passing scores being considered by the examiner. Using the data (x_i, y_i), it is possible to compute the sample proportion of examinees, say $\hat{p}_{1j}$, for whom $x_i \geq x_{0j}$ and y_i=0. Thus, $\hat{p}_{1j}$ estimates p_{1j}, the probability of a false-positive decision using the passing score x_{0j}. In a similar manner we can estimate p_{0j}, the probability of a false-negative decision, with $\hat{p}_{0j}$. Note that both $\hat{p}_{1j}$ and $\hat{p}_{0j}$ are unbiased and maximum likelihood estimators. Following Huynh (1976, 1977) as well as Novick and Lewis (1974), we consider the case of constant losses for the two possible error types. Thus, we let $c_s > 0$ represent the cost of predicting success when the examinee fails on the external criterion. Correspondingly, we let $c_f > 0$ denote the cost of predicting failure when the reverse decision should have been made. The risk or expected loss associated with the passing score x_{0j} is

$$r_j = c_s p_{1j} + c_f p_{0j}.$$

Let $\hat{r}_j = c_s\hat{p}_{1j} + c_f\hat{p}_{0j}$ denote the unbiased estimate of r_j and let $\hat{r}_{[1]} \leq \hat{r}_{[2]} \leq \ldots \leq \hat{r}_{[k]}$ represent the ordered values of the $\hat{r}_j$'s. Since $\hat{p}_{1j}$ and $\hat{p}_{0j}$ are maximum likelihood estimates of p_{1j} and p_{0j} it follows from a theorem given by Zehna (1966) that $\hat{r}_j$ is a maximum likelihood estimate of r_j. Since the optimal passing score is the one for which the risk is smallest, the natural choice for a passing score is the one associated with $\hat{r}_{[1]}$.

A similar procedure for determining an optimal decision rule for predicting success or failure on an external criterion is described by Johns (1961). In contrast to the present situation, he uses a more complicated decision rule for predicting success or failure on the external criterion. (See also, Ott and Kronmal, 1976; Aitchison and Aitken, 1976.)

3. A Lower Bound to the Probability of Choosing an Optimal Passing Score

The procedure outlined above seems to be the "obvious" and most "natural" method of determining an optimal passing score when there is an external criterion. We now consider the problem of determining N so that the probability of correctly identifying an optimal passing score is reasonably high, say greater than or equal to P^*. If the passing score associated with $r_{[1]}$ is chosen, we say that a correct decision (CD) has been made. Symbolically we want to find N so that

(3.1) $\quad P(CD) \geq P^*$.

The results reported here can also be used to determine a lower bound on the probability of a correct decision for a given N. For a variation of the present solution which does not yield an exact value for N see Gibbons, et al., (1977, p. 172).

Let $r_{[2]} - r_{[1]} = \delta$. To solve this problem the investigator specifies the smallest value of δ, say δ^*, that is worth detecting. That is, if $r_{[2]} - r_{[1]} < \delta^*$, we are indifferent as to whether the passing score associated with $r_{[2]}$ or the one associated with $r_{[1]}$ is used. If, however, $r_{[2]} - r_{[1]} \geq \delta^*$, we want to be reasonably certain of choosing the passing score associated with $r_{[1]}$. If $\delta^* = 0$, it may be impossible to choose N so that (3.1) is satisfied.

The choice for the constants P^* and δ^* is not an easy task. Rather than discuss this problem here, the reader is referred to Gibbons, Olkin and Sobel (1977, Chapter 1). (See, in particular, pp. 19-21.)

The Case $c_s = c_f = 1$, $k = 2$

Before considering the general case $c_s > 0$, $c_f > 0$ and $k \geq 2$, some exact results are given for the special case $c_s = c_f = 1$ and $k = 2$. Let $x_{01} < x_{02}$ denote the two passing scores under consideration. A randomly chosen examinee having an observed score x falls into one of six mutually exclusive categories as depicted in Table 1. The cell entries in Table 1 represent the probabilities associated with these events. t_{11}, for example, is the probability that a randomly chosen examinee has an observed score greater than or equal to x_{02} and is successful on the external criterion (y=1). The risk using the passing score x_{02} is $t_{10} + t_{21} + t_{31}$ and the risk using x_{01} is $t_{10} + t_{20} + t_{31}$.

Thus, the choice between the two passing scores x_{01} and x_{02} reduces to determining whether t_{21} is larger or smaller than t_{20}. Let w_0 and w_1 be the number of examinees falling in the cells corresponding to t_{20} and t_{21}, respectively. Consistent with the decision rule for determining the optimal passing score, the decision $t_{21} > t_{20}$ is made if $w_1 > w_0$. If $w_0 < w_1$, the reverse is said to be true. If $w_0 = w_1$, we decide at random which of the two cells has the larger probability.

Without loss of generality suppose $t_{20} < t_{21}$ and note that $\delta = r_{[2]} - r_{[1]} = t_{21} - t_{20}$. Let $u = \min(w_1 - 1, N - w_1)$. The probability of a correct decision is

$$(3.2) \qquad P(w_0 < w_1) + \tfrac{1}{2} P(w_0 = w_1) = \sum_{w_1=0}^{N} \sum_{w_0=0}^{u} P(w_0, w_1) + \tfrac{1}{2} \sum_{w=0}^{N} P(w, w).$$

Since the conditional distribution of w_0 given w_1 is binomial with parameters $N - w_1$ and $(t_{21} - \delta)/(1 - t_{21})$, the right-hand side of (3.2) may be written as

$$(3.3) \qquad \sum_{w_1=0}^{N} \left[\tfrac{1}{2} \binom{N-w_1}{w_1} \left(\frac{t_{21} - \delta}{1 - t_{21}} \right)^{w_1} \left(\frac{1 - 2t_{21} + \delta}{1 - t_{21}} \right)^{N-2w_1} + \sum_{w_0=0}^{u} \binom{N-w_0}{w_0} \left(\frac{t_{21} - \delta}{1 - t_{21}} \right)^{w_0} \left(\frac{1 - 2t_{21} + \delta}{1 - t_{21}} \right)^{N-w_1-w_0} \right] \binom{N}{w_1} t_{21}^{w_1} (1 - t_{21})^{N-w_1}$$

where the binomial coefficient $\binom{N-w_1}{w_1}$ is taken to be zero when $w_1 > N/2$. From Sobel and Huyett (1957, p. 565, Lemma 1) it can be verified that for fixed t_{21} the expression inside the brackets of expression (3.3) decreases with decreasing values of δ. Since the expression outside the brackets does not depend on δ, it follows that the minimum probability of a correct decision occurs at $\delta = \delta^*$. Thus, N is sufficiently large to guarantee (3.1) if the

value of (3.2) is greater than or equal to P^* for $\delta = \delta^*$ and all admissible values of t_{21}.

Let t' be the value of t_{21} that minimizes (3.4). The value of t' can be determined using numerical techniques. If for a given N, the value of (3.4) evaluated at t' is greater than or equal to P^*, it is evident that (3.1) is guaranteed regardless of the actual value of t_{21}. Thus, we can find the minimum N so that (3.1) holds.

It is of interest to note that when numerical procedures were used to minimize (3.4), the minimum always occurred at $t_{21}=(1+\delta^*)/2$. No proof is given that the minimum is always attained at this point but it is readily verified that this is true in the asymptotic case. To see this, observe that for N large and $t_{20}+\delta^*=t_{21}$, $w=w_0-w_1$ is approximately normal with mean $-\delta^*$ and variance $N[(t_{21}-\delta^*)(1-t_{21}+\delta^*) + t_{21}(1-t_{21}) + 2(t_{21}-\delta^*)t_{21}]$. Moreover, $P(CD)=P(w<0)$ which decreases as the variance of w gets large. Since the variance of w attains its largest admissible value at $t_{21}=(1+\delta^*)/2$, the desired result follows.

An Approximate Solution for Determining N

When $t_{21} = (1+\delta^*)/2$ and $t_{20}=t_{21}-\delta^*$, the problem of comparing t_{20} to t_{21} reduces to the problem of determing whether t_{21} is above or below 1/2. Via the central limit theorem, it can be seen that the smallest N so that (3.2) is greater than or equal to P^* is given approximately by

$$N = (\lambda/\delta^*)^2 \tag{3.4}$$

where λ is the P^* quantile of the standard normal distribution.

Table 2 gives the N values using the two procedures just described. The first entry in the cell corresponding to a pair of values of P^* and δ^* gives the smallest value of N so that (3.2) is greater than or equal to P^* with $t_{21}=(1+\delta^*)/2$. The second entry gives the value of N using (3.4). For example, if $P^*=.9$, $\delta^*=.2$, the value of N using (3.2) or (3.4) is 41. Thus, the probability of selecting the best of two passing scores when $c_s=c_f=1$ is at least .9 when 41 examinees are randomly sampled and when $\delta^*=.2$.

Solution for any Positive c_s, c_f with k=2

A brief outline is given on how the solution for $c_s= c_f=1$ and k=2 may be extended to the case where c_s and c_f are any two positive numbers. First observe that when comparing r_1 to r_2 we can, for convenience, have one of the values of c_s or c_f set equal to 1. Suppose, for example, $c_s=1$. In terms of the parameters t_{ij}, we are comparing $t_{10}+c_f(t_{21}+t_{31})$ to $t_{10}+t_{20}+c_f t_{31}$. For the case $c_f t_{21}-t_{20}=\delta\geq\delta^*$, the probability of a correct decision is given by expression (3.2) or (3.3) after an appropriate adjustment is made in the limits of the summations. For the case $t_{20}-c_f t_{21}=\delta\geq\delta^*$, a similar expression can be obtained for the probability of a correct decision. In both cases (i.e., $t_{20}>c_f t_{21}$ and $t_{20}<c_f t_{21}$) the probability of a correct decision is minimized when $\delta=\delta^*$. When $\delta=\delta^*$, the expression for the probability of a correct decision reduces to a function of either t_{21} or t_{20}. Moreover, it is possible to minimize

this function in the manner described above which means that we are able to find the smallest N (or close to it) so that (3.1) is guaranteed.

The case k > 2

This last section considers the more general case in which there are $k \geq 2$ passing scores from which to choose. Let $T^* = (1-P^*)/(k-1)$ and let A_{1i} represent the event of deciding $r_{[1]} < r_{[i]}$. The requirement (3.1) becomes

$$P(\bigcap_{i=2}^{k} A_{1i}) \geq P^*.$$

From the Bonferroni inequalities (Feller, 1968, p. 110)

$$P(CD) \geq 1 - \sum_{i=2}^{k} P(A_{1i}^c).$$

From the results given above we can determine the smallest N so that for any i, $P(A_{1i}) \geq 1-T^*$ which implies that $P(A_{1i}^c) \leq T^*$. Thus,

$$\sum_{i=2}^{k} P(A_{1i}) \leq (k-1)T^*$$

which means that

$$P(CD) \geq 1-(k-1)T^*$$
$$= P^*$$

as was required.

TABLE 1

Probabilities Associated with Two Passing Scores

	y=1	y=0
$x \geq x_{02}$	t_{11}	t_{10}
$x_{02} > x \geq x_{01}$	t_{21}	t_{20}
$x < x_{01}$	t_{31}	t_{30}

TABLE 2

Values of $\hat{N}$ Using (3.2) and (3.4)

δ^*	p^* .75	.9	.95	.975	.99
.1	45	163	269	381	536
	46	163	272	384	543
.2	11	41	67	95	133
	12	41	68	96	136

REFERENCES

Aitchison, J. and Aitken, C. G. G. Multivariate binary discrimination by the kernel method. Biometrika, 1976, 63, 413-420.

Feller, W. An introduction to probability theory and its applications. Vol. I, New York: John Wiley, 1968.

Gibbons, J., Olkin, I., and Sobel, M. Selecting and ordering populations: A new statistical methodology. New York: John Wiley, 1977.

Griffiths, D. A. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution on the total number of cases of a disease. Biometrics, 1973, 29, 637-648.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

Huynh, H. Two simple classes of mastery scores based on the beta-binomial model. Psychometrika, 1977, 42, 601-608.

Johns, M. V. An empirical Bayes approach to non-parametric two-way classification. In H. Solomon (Ed.) Studies in item analysis and prediction, Stanford, CA: Stanford University Press, 1961.

Keats, J. A. and Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Meredith, W. and Kearns, J. Empirical Bayes point estimates of latent trait scores without knowledge of the trait distribution. Psychometrika, 1973, 38, 533-554.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Novick, M. R. and Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, and W. James Popham (Eds.) Problems in criterion-referenced measurement. CSE Monograph No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Olkin, I. Monotonicity properties of Dirichlet integrals with applications to the multinomial distribution and the analysis of variance. Biometrika, 1972, 59, 303-307.

Ott, J. and Kronmal, R. A. Some classification procedures for multivariate binary data using orthogonal functions. Journal of the American Statistical Association, 1976, 71, 391-399.

Sobel, M. and Huyett, M. Selecting the best one of several binomial populations. Bell System Technical Journal, 1957, 36, 537-576.

Van der Linden, W. J. and Mellenbergh, G. J. Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1977, 1, 593-599.

Wilcox, R. R. Estimating the parameters of the beta-binomial distribution. Educational and Psychological Measurement, 1979, in press.

Zehna, P. W. Invariance of maximum likelihood estimation. Annals of Mathematical Statistics, 1966, 37, 744.

ESTIMATING THE PARAMETERS OF THE BETA-BINOMIAL DISTRIBUTION

Rand R. Wilcox
Center for the Study of Evaluation
University of California at Los Angeles

ABSTRACT

For some situations the beta-binomial distribution might be used to describe the marginal distribution of test scores for a particular population of examinees. To use this distribution it is necessary to estimate two parameters which characterize the beta-binomial model. One method of estimating these parameters is to approximate the maximum likelihood estimates using some appropriately chosen iterative technique. When the number of examinees is small, however, it is not clear that this method of estimation is justified since the iterative approximations might not converge to the maximum likelihood estimate. Using Monte Carlo techniques, this paper compares the Newton-Raphson approximation to the maximum likelihood estimate with several other possible procedures. It is found that the Newton-Raphson method should be used when it yields admissible (positive) results.

For certain types of tests for which an item sampling model applies and when the items are scored dichotomously, the beta-binomial distribution (BBD), also known as the negative hypergeometric or Polya distribution, may be useful when describing the marginal distribution of observed scores for a particular population of examinees (Keats and Lord, 1962; Lord, 1965; Lord and Novick, 1968, Chapter 23). Wilcox (1977) and Huynh (1976), for example, have used this distribution to describe a mastery test.

The BBD arises as follows: Consider the binomial probability function

$$(1) \qquad f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}$$

If we view p as arising from a beta distribution g(p) with parameters r and s, the marginal distribution of x is

$$(2) \qquad f(x) = \int \binom{n}{x} p\,(1-p)^{n-x}\, g(p)dp$$

$$= \binom{n}{x} \frac{B(r+x,\ n-x+s)}{B(r,\ s)}$$

where B is the usual beta function. In mental test theory p is sometimes referred to as the percent correct true score of an examinee and g(p) is the distribution of true scores over a population of examinees. Various properties and extensions of (2) are given by Ishii and Hayakawa (1960).

A fundamental problem when using the BBD is estimation of the parameters r and s. Two methods are typically employed -- an approximation to the maximum likelihood estimate (MLE) and the method of moments. Results reported by Shenton (1950) suggest that the two estimation procedures will yield nearly the same results when the number of examinees is

large. As indicated below, however, this may not always be the case, particularly when k is small.

The purpose of this paper is to examine the difficulties which might occur when using the method of moments or an approximation to the MLE of the parameters of the BBD. We also compare these two estimation techniques with several other procedures which might be used.

THE METHOD OF MOMENTS

Computationally, the easiest method to use when estimating r and s is the method of moments. This method simply expresses r and s in terms of the first and second moments of the BBD. Since the first two moments of the BBD can be estimated, we obtain an estimate of r and s.

More specifically, let $x_1,\ldots,x_k$ be a random sample of size k from a BBD with parameters r and s. From Lord and Novick, (1968, Chapter 23) it can be verified that

$$\hat{r}_1 = \hat{\gamma}_1^2(1-\hat{\gamma}_1)/\hat{\gamma}_2^2 - \hat{\gamma}_1$$

$$\hat{s}_1 = \hat{\gamma}_1\hat{\gamma}_2^{-2}\,(1-\hat{\gamma}_1)^2 + \hat{\gamma}_1 - 1$$

may be used as estimates of r and s, where

$$\hat{\gamma}_1 = (kn)^{-1} \sum_{i=1}^{k} x_i$$

$$\hat{\gamma}_2 = \hat{\gamma}_3 - \hat{\gamma}_1^2$$

$$\hat{\gamma}_3 = [kn(n-1)]^{-1} \sum_{i=1}^{n} (x_i^2 - x_i).$$

From Keats and Lord (1962) we see that the moment estimates may be written as

$$\hat{r}_1 = (-1+\alpha_{21}^{-1})\hat{\mu}$$

$$\hat{s}_1 = -\hat{r}_1+n/\alpha_{21}-n$$

where α_{21} is the KR21 reliability coefficient and $\hat{\mu}$ is the sample mean of the observed data.

MAXIMUM LIKELIHOOD ESTIMATES

Skellam (1948) describes an iterative procedure for obtaining maximum likelihood estimates of r and s using the digamma function. More recently, Griffiths (1973) has shown that the digamma function can be avoided. Instead, one determines the value of π and θ, say $\hat{\pi}$ and $\hat{\theta}$, so that

$$(3) \qquad \sum_{i=0}^{n-1} \frac{S_n-S_i}{\pi+i\theta} - \frac{S_{n-1-i}}{1-\pi+i\theta} = 0$$

$$(4) \qquad \sum_{i=1}^{n-1} i \left\{\frac{S_n-S_i}{\pi+i\theta} + \frac{S_{n-1-i}}{1-\pi+i\theta} - \frac{S_n}{1+i\theta}\right\} = 0$$

where $S_i = \sum_{x=0}^{i} f_x$ and where f_x is the number of examinees obtaining an observed x.

The estimates of r and s are

$$(5) \qquad \hat{r}_2 = \hat{\pi}\hat{\theta}^{-1}$$

$$(6) \qquad \hat{s}_2 = \hat{\theta}^{-1} - \hat{\pi}\hat{\theta}^{-1}$$

Solutions to (3) and (4) can be obtained iteratively using the Newton-Raphson method. The method of moments gives a good starting point.

Maximum likelihood estimation is a popular and well-known procedure. Wolfowitz (1965) gives an interesting discussion of its asymptotic optimal properties. However, for small sample sizes, there may be technical difficulties when using $\hat{r}_2$ and $\hat{s}_2$. Kale (1962) discusses the use of several iterative solutions which approximate a maximum likelihood estimate. He concludes that the Newton-Raphson procedure is justified for large samples since the iterative solution approaches the MLE with probability one when the initial estimate is consistent. Our own experience when the number of examinees is small is that the Newton-Raphson procedure will sometimes converge to negative values for r and s. Since the BBD is defined only for positive values of r and s, negative values can hardly be interpreted as MLE's.

Barnett (1966) points out that difficulties with iterative approximations to MLE's can occur even when the regularity conditions assumed by Kale are satisfied. In particular, there may be multiple roots to the likelihood equation even when a unique consistent root exists.

We see that there is some doubt about when the Newton-Raphson procedure should be applied when estimating r and s. Accordingly we intend to deal with this problem at least partially by considering the case where the number of examinees is small. Before doing so, we first describe three alternatives to the estimation techniques described above. Two of these procedures have certain optimal properties in the asymptotic case. The other procedure is based on bias reduction techniques.

ALTERNATIVE ESTIMATION PROCEDURES

Our first alternative estimation technique is to iterate only once with the Newton-Raphson procedure using the moment estimates $\hat{r}_1$ and $\hat{s}_1$ as the initial values. Following Griffiths (1973) we can write this estimate more formally as follows: Consistent with our earlier notation we let

CSE
ARCHIVE COPY

$\pi = r(r+s)^{-1}$ and $\theta = (r+s)^{-1}$. The log likelihood for the BBD is given by

$$\ln L = c - S_n \sum_{i=1}^{n-1} \ln(1+i\theta) + \sum_{i=0}^{n-1} \{(S_n - S_i)\ln(\pi+i\theta) + S_{n-1-i} \ln(1-\pi+i\theta)\}$$

where the constant c becomes zero after differentiation and so plays no role in the estimation of r and s, as will become apparent. Define the vector G as

$$G = \left(\frac{\partial \ln L}{\partial r}, \frac{\partial \ln L}{\partial s}\right)$$

and the 2x2 matrix Λ as

$$\Lambda = \begin{pmatrix} \frac{\partial^2 \ln L}{\partial^2 r} & \frac{\partial^2 \ln L}{\partial r\, \partial s} \\ \frac{\partial^2 \ln L}{\partial s\, \partial r} & \frac{\partial^2 \ln L}{\partial^2 s} \end{pmatrix}$$

and consider

$$(\hat{r}_3, \hat{s}_3) = (\hat{r}_1, \hat{s}_1) - G\Lambda^{-1} \tag{7}$$

The statistics $\hat{r}_3$ and $\hat{s}_3$ are simply the values obtained after one iteration using the Newton-Raphson procedure.

We verify that (7) is BAN (best asymptotically normal). As noted by Blischke (1964, p. 522) it is sufficient to verify that $\hat{r}_1$ and $\hat{s}_1$ are $o(\sqrt{k})$-consistent. That is, we must verify that $k^{\frac{1}{2}}(\hat{r}_1 - r)$ and $k^{\frac{1}{2}}(\hat{s}_1 - s)$ are bounded in probability uniformly in k. To show that $\hat{r}_1$ and $\hat{s}_1$ are $o(\sqrt{k})$-consistent we need only show that they are asymptotically normal (Ferguson, 1958). Applying a theorem given by Cramér (1946, p. 366) $\hat{r}_1$ and $\hat{s}_1$ are asymptotically normal and so (7) is a BAN estimate of r and s.

For notational convenience let $t = (r, s)$. The fourth estimation procedure arises as follows: Let

$$\chi^2(t) = k \sum_{x=0}^{n} \frac{\left[f_x - f(x)\right]^2}{f(x)}$$

The first three terms in the Taylor series expansion of $\chi^2(t)$ about $\hat{t} = (\hat{r}_1, \hat{s}_1)$ are

$$\chi^2(t) = \chi^2(\hat{t}) + \dot{\chi}^2(\hat{t})' (t-\hat{t}) + \tfrac{1}{2}(t-\hat{t})' \ddot{\chi}^2(t)(t-t) \tag{8}$$

where

$$\dot{\chi}^2(t)' = \left(\frac{\partial}{\partial r} \chi^2(t), \frac{\partial}{\partial s} \chi^2(t)\right)$$

$$\ddot{\chi}^2(t) = \begin{pmatrix} \frac{\partial^2}{\partial r^2} \chi^2(t) & \frac{\partial^2}{\partial r\, \partial s} \chi^2(t) \\ \frac{\partial^2}{\partial s\, \partial r} \chi^2(t) & \frac{\partial^2}{\partial^2 s^2} \chi^2(t) \end{pmatrix}$$

Since $\hat{t}$ is $O(\sqrt{k})$-consistent, the value of t which minimizes (8) is BAN (Ferguson, 1958). This value is

$$(\hat{r}_4, \hat{s}_4) = \hat{t} - \ddot{\chi}^2(\hat{t})^{-1} \dot{\chi}^2(\hat{t}) \tag{9}$$

Expressions for $\dot{\chi}$ and $\ddot{\chi}$ in terms of the gamma function can be derived in a straightforward manner. To evaluate the derivative of the gamma function we use some results given by Johnson and Kotz (1969, p. 7). In particular we have that

$$\Gamma'(r) = \Gamma(r)\, \psi(r)$$

where ψ is the digamma function. For $r \geq 2$ we use the approximation

(10) $\psi(r) \doteq \ln(r-\frac{1}{2})$.

For $0 < r < 2$ we use

(11) $\psi(r) \doteq \psi(r+2) - r^{-1} - (r+1)^{-1}$

where ψ on the right-hand side of (11) is given by (10). Proceeding in a similar manner we have that

$$\Gamma''(r) = \Gamma(r)\psi'(r) + \Gamma'(r)\,\psi(r)$$

For ψ' we use the approximation

$$\psi'(r) \doteq (r-\tfrac{1}{2})^{-1}, \quad r \geq 2$$

$$\doteq \psi(r+2) + r^{-2} + (r+1)^{-2}, \quad 0 < r < 2.$$

The fifth estimation procedure is the jacknife (Quenouille, 1956) using one-at-a-time omission. The purpose of this estimator is to reduce the bias in the moment estimators $\hat{r}_1$ and $\hat{s}_1$. Miller (1974) gives a good review of this approach to estimation.

Let $\tilde{r}_i$ be the moment estimate of r with the ith observation removed. That is, $\tilde{r}_i$ is computed in the same manner as $\hat{r}_1$ using the observations $x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k$. The statistic $\tilde{s}_i$ is defined in an analogous fashion. The jacknife estimate of r is

$$\hat{r}_5 = k^{-1} \sum_{i=1}^{k} T_i$$

where

$$T_i = k\hat{r}_1 - (k-1)\tilde{r}_i$$

EMPIRICAL RESULTS COMPARING THE FIVE ESTIMATION PROCEDURES

As indicated earlier, the estimates of r and s using any of the five estimation procedures described above might be negative. When this occurs it may correspond to a negative estimate of the variance of the assumed beta prior. In this case it seems reasonable to follow Wilcox (1977) and say that the BBD degenerates to a binomial distribution. This corresponds to letting $r \to \infty$, $s \to \infty$ with $r(r+s)^{-1}$ a constant. If, however, we allow our estimate of r and s to have an infinite value and if we use squared error loss, then the risk of all five estimators is also infinite and so comparisions become meaningless. To circumvent this problem we consider two aspects of the statistics described above. Using Monte Carlo techniques we estimate for each $\hat{r}_i$ and $\hat{s}_i$ (i=1...,5) the probability of an admissible (positive) estimate of r and s. For these admissible estimates we also compute $E(r-\hat{r}_i)^2$ and $E(s-\hat{s}_i)^2$. In otherwords, we compute expected squared error loss by ignoring inadmissible estimates of the parameters involved. In this manner we will have some indication about how any two estimators compare when they both yield admissible results.

Table 1 gives the results of some of our Monte Carlo experiments for n=10, k=10. We also considered n=5, 20; k=25. We decided not to report these results however, since they led to the same conclusions as those based on the results in Table 1. All estimates in Table 1 are based on 1000 iterations. The maximum number of allowed iterations in the Newton-Raphson procedure was 20. This limit was imposed so as to avoid excessive

computer costs. In practice the Newton-Raphson procedure usually converges after only a few iterations, at least for a tolerance error of 10^{-4}. In all of our empirical studies we considered only situations where both r and s are less than 20. Experience with the BBD suggests that r and s will usually be in this range. (See, e.g., Shenton, 1950; Ishii and Hayakawa, 1960; Williams, 1975).[1]

Column one in Table 1 indicates the estimation procedure used. Thus, the results for estimator 2 are the results obtained when using $\hat{r}_2$ and $\hat{s}_2$. Consider, the example, the case r=s=.5. Table 1 indicates that $E(\hat{r}_2-r)^2=1.25$ and $E(\hat{s}_2-s)^2=.377$.

We decided to report the results for the values of r and s shown in Table 1 primarily because this demonstrates that the accuracy of all five estimation procedures might be poor when the number of examinees is small. In terms of comparing the various estimation techniques all combinations of r and s gave the same general results. In particular, we found that among the admissible estimates of r and s, the Newton-Raphson approximation to the MLE was most accurate of all. As can be seen, the improvement of $\hat{r}_2$ over the other four estimators, in terms of squared error loss, is often quite dramatic. A similar result was found for $\hat{s}_2$. Thus, to be conservative, it appears that an investigator should always compute an approximation to the MLE of r and s when the beta-binomial model is being used.

As previously indicated, experience suggests that $\hat{r}_1$ and $\hat{r}_2$ will usually give nearly the same results when the number of examinees is large. To find out whether this was the case for the situations considered here, we examined the 1,000 values of $\hat{r}_1$ and $\hat{s}_2$ that resulted from the 1,000 iterations in a

[1] Ishii and Hayakawa also report a case where both r and s are close to 100. Consideration of cases where r and s are large would not change the conclusions made here.

Monte Carlo experiment. This was done for several of the Monte Carlo experiments that were conducted. Usually the values of $\hat{r}_1$ and $\hat{r}_2$ were nearly the same. On occasion, however, $\hat{r}_1$ yielded a wildly inaccurate estimate of r, while $\hat{r}_2$ did not, which accounts for the large difference in expected squared error loss. The same is also true for $\hat{s}_1$ and $\hat{s}_2$.

As a final comment, we note that $\hat{r}_2/(\hat{r}_2+\hat{s}_2)$ may be used to approximate the MLE of $\mu=r/(r+s)$, the mean of the distribution of true scores. However, our results do not necessarily imply that this estimate should be used. Kleinman (1973) shows that $\hat{\mu}=(kn)^{-1}\sum_{i=1}^{k} x_i$ is a highly efficient estimate of μ. In fact Kleinman gives several reasons why one might prefer $\hat{\mu}$ to $\hat{r}_2(\hat{r}_2+\hat{s}_2)$ when estimating μ.

TABLE 1

Estimator	Probability of an admissible estimate of r	Probability of an admissible estimate of s	$E(\hat{r}-r)^2$	$E(\hat{s}-s)^2$
		$(r, s) = (.5, .5)$		
1	99.9	99.9	.45	.96
2	82.1	83.0	1.25	.377
3	96.3	87.8	.75	1.30
4	93.9	94.5	28.1	95.3
5	99.9	99.9	.46	.962
		$(r, s) = (2, 2)$		
1	97.4	97.4	1.5×10^4	1.5×10^4
2	77.5	79.1	3.5	3.4
3	90.1	92.8	23.6	24.5
4	66.5	64.8	2.2×10^6	2.2×10^6
5	97.4	97.4	1.5×10^4	1.5×10^4
		$(r, s) = (5, 5)$		
1	81.4	81.4	435.0	429.6
2	86.3	80.4	24.8	24.8
3	98.0	98.3	5.7×10^3	1.0×10^4
4	52.9	53.1	2.7×10^5	2.6×10^5
5	81.4	81.4	435.1	429.6
		$(r, s) = (2, 9)$		
1	80.3	80.3	9.2×10^{10}	1.5×10^{12}
2	86.1	79.8	3.33	65.7
3	79.9	80.3	247.8	5.3×10^3
4	52.8	51.7	1.4×10^{13}	2.3×10^{14}
5	80.3	80.3	9.2×10^{10}	1.5×10^{12}

REFERENCES

Barnett, V. D. Evaluation of the maximum - likelihood estimator where the likelihood equation has multiple roots. Biometrika, 1966, 53, 151-165.

Blischke, W. R. Estimating the parameters of mixtures of binomial distributions. Journal of the American Statistical Association, 1964, 59, 510-528.

Cramér, H. Mathematical Methods of Statistics. Princeton, New Jersey: Princeton University Press, 1946.

Ferguson, T. S. A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. Annals of Mathematical Statistics, 1958, 29, 1046-1062.

Griffiths, D. A. Maximum likelihood estimation for the beta - binomial distribution and an application to the household distribution on the total number of cases of a disease. Biometrics, 1973, 29, 637-648.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

Ishii, G. and Hayakawa, R. On the compound binomial distribution. Annals of the Institute of Statistical Mathematics, 1960, 12, 69-80.

Johnson, N. and Kotz, S. Discrete Distributions. New York: John Wiley, 1969.

Kale, B. K. On the solution of likelihood equations by iteration processes. The multiparametric case. Biometrika, 1962, 49, 479-486.

Keats, J. A. and Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Kleinman, J. C. Proportions with extraneous variance: single and independent samples. Journal of the American Statistical Association, 1973, 68, 46-53.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass: Addison-Wesley, 1968.

Miller, R. G. The jacknife - a review. Biometrika, 1974, 61, 1-15.

Quenouille, M. H. Notes on bias in estimation. Biometrika, 1956, 43, 353-360.

Shenton, L. R. Maximum likelihood and the efficiency of the method of moments. Biometrika, 1950, 37, 111-116.

Skellam, J. G. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. Journal of the Royal Statistical Society, 1948, 10B, 257-261.

Wilcox, R. R. Estimating the likelihood of a false-positive or false-negative decision with a mastery test: an empirical Bayes approach. Journal of Educational Statistics 1977, 2, 289-307.

Williams, D. A. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics, 1975, 31, 949-952.

Wolfowitz, J. Asymptotic efficiency of the maximum likelihood estimator. Teoriia Veroiatnostei, 1965, 10, 267-281.